Intelligent and Distributed Data Warehouse for Student´s Academic Performance Analysis

Jesús Silva[1], Lissette Hernández[2], Noel Varela[3], Omar Bonerge Pineda Lezama[4], Jorge Tafur Cabrera[5], Bellanith Ruth Lucena León Castro[6], Osman Redondo Bilbao[7], and Leidy Pérez Coronel – Hernandez[8]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
jesussilvaUPC@gmail.com
[2,3] Universidad de la Costa, St. 58 #66, Barranquilla, Atlántico, Colombia
{lhernand31, nvarela2}@cuc.edu.co
[4] Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras
omarpineda@unitec.edu
[5,6,7,8] Corporación Universitaria Latinoamericana, Barranquilla, Colombia.
{jtafur, brleonc, oredondo}@ul.edu.co; leidyperezcoronel89@gmail.com

Abstract. In the academic world, a large amount of data is handled each day, ranging from student´s assessments to their socio-economic data. In order to analyze this historical information, an interesting alternative is to implement a Data Warehouse. However, Data Warehouses are not able to perform predictive analysis by themselves, so machine intelligence techniques can be used for sorting, grouping, and predicting based on historical information to improve the analysis quality. This work describes a Data Warehouse architecture to carry out an academic performance analysis of students.

Keywords: Intelligent data retrieval; data warehouse; Unique Identification Number; Academic Performance.

1. Introduction

One of the most commonly used actions in educational institutions to give value to information and to support decision-making processes is the design of reports. The report designing is an exploratory action where certain crosses of data are made and, depending on the results, other criteria are analyzed until reaching a point in which the results are enough to make decisions about the organization. Support for the decision-making process can be provided by specially designed systems such as [1] DSS (Decision Support Systems), which can generate configurable reports on a regular, quick, and easy basis, as expressed in [2].

On the other hand, Data Warehouses (DW) are electronic data repositories specially designed for generating reports and data analyses [3], [4]. The distinctive features of DW about systems described above are the following: (i) they are flexible, (ii) integrate all points of interest about the organization, (iii) can efficiently handle large amounts of data, and (iv) allow the creation and calculation of management indicators. In addition, the DW are designed with the aim to be efficient in the

analysis requirements for strategic levels in organizations, directly considering the organizational strategic objectives [5]. In the same way, DW let efficiently analyze historical information and allow to visualize trends in the behavior of management indicators over time. However, even though the historical information can provide an indication of the historical trend followed by an indicator, it is not enough to certainly predict any particular indicator. A DW can certainly provide a solid basis for analysis and initial performance in the Machine Intelligence techniques [6] that allow learn the patterns of these indicators to predict future patterns. For the latter, the Artificial Neural Networks (ANN) are algorithms that can associate or classify patterns, compress data, control processes and approximate nonlinear functions [7], [8].

In this work, the DW has been designed for the behavior analysis of approval and advance in a curriculum with real data of the curricula vitae of students from different universities in Colombia that offer the Industrial Engineering career. The DW is not focused only on the analysis of historical behaviors of students, but also has been thought of as a base architecture for the prediction of future trends through ANN techniques. This research proposes the approach of data retrieval from an Intelligent Distributed Data Warehouse (IDDW), which is a hierarchical distributed data store of N levels.

2. Theoretical Review

2.1 Artificial Neural Networks

Artificial Neural Networks (ANN) can learn from data and can be used to construct reasonable input-output mapping, with no prior assumptions on the statistical model of the input data (Haykin, 2009) [9]. ANN have non-linear modeling capability with a data-driven approach so that the model is adaptively formed based on the features presented from the data (Zhang 2003) [10]. An introduction to ANN model specifications and implementation and their approximation properties has been provided from an econometric perspective (Kuan 2008) [11]. Several studies show that ANN can solve a variety of challenging computational problems, such as pattern classification, clustering or categorization, function approximation, prediction or forecasting, optimization (traveling salesman problem), retrieval by content, and control (Jain, Mao, and Mohiuddin, 1996) [12].

Some studies of ANN application related to financial early warning models have been conducted by Sevim et al. (2014) [13], as well as Sekmen and Kurkcu (2014) [14] who used ANN as a classifier with a categorical output. Other authors used ANN as financial forecasting models with continuous value. Some of them are Singhal and Swarup (2011) [15], as well as Mombeini and Yazdani-Chamzini (2015) [16] who implemented ANN with a single-step prediction output. A previous study on ANN forecasting model was also proposed by Kulkarni and Haidar (2009) [17] for a multi-step prediction with a direct strategy, so the number of models is equal to the number of the prediction horizon. In the context of basic commodity prices, the need for prediction is not limited to one-step forward but could be extended to include multi-step ahead predictions. Three strategies to tackle the multi-step forecasting problem

can be considered, namely recursive, direct, and multiple output strategies (Bontempi, Ben Taieb, and Borgne 2013) [18]. The Multiple Input Multiple Output (MIMO) techniques train a single prediction model f that produces vector outputs of future prediction values. The study proposes to Multi-Layer Perceptron with Multiple Input and Multiple Output (MLP-MIMO) as an agricultural product price prediction model coupled with the variation coefficient from the Colombian state price reference to the criteria of warning level.

2.2 Data Retrieval

Initially, the data retrieval techniques were restricted only to the centralized processing, as discussed by Duan L. et al. (2009) [19]. But, according to Abhay K. et al. (2015) [20], the data retrieval from the distributed data warehouse refers to the implementation of the classic procedure for retrieving data in a distributed computing environment that seeks to maximize the use of available resources (communication networks, computers, and databases). Some algorithms and systems used for the distributed retrieval of databases are the following: the partition algorithm of Savasere A. et al. (1995) [21]; Multiagent system based on JAVA JAM by Stolfo S. et al. (1997) [22] and Prodromidis A. et al. (2000) [23]; Parthasarathy S. et al. (2000) [24] in D-DOALL uses the primitive distributed do-all to easily program the task of independent retrieval in a workstations network; Grossman R. L. et al. (1999) [25] proposed the Papyrus, a JAVA-based system which aims to wide-area distributed data on clusters and meta-clusters; and the system based on Java for distributed enterprises by Chattratichat J. et al. (1999) [26].

The data retrieval in a highly parallel environment on multiple processors was explained by Wang L. et.al. (2013) [27]. There are two commonly used parallel programming models: Subprocesses (POSIX subprocesses by Butenhof D. R. (1997) [28]) and message passing (OpenMP by Duan L. et al. (2009)) [19]. Modern programming languages are also structured to efficiently use innovative architectures. There are parallel programming paradigms focusing on parallelizing the algorithms on multiprocessor systems and networks. OPENMP and MPI are used to achieve the parallelization of shared and distributed memory. CUDA is a programming language that is designed for parallel programming used by Garciarena U. et al. (2015) [29]. In CUDA, the threads access different memories of the GPU. CUDA offers a model of data parallel programming which is incomplete without discussing the more recent approach called MapReduce that can process large amounts of data in a highly parallel way, as shown by Bhaduri K. et al. (2008) [30]. Several data recovery algorithms have been modified for parallel processing architectures as discussed in Parthasarathy S. et al. (2000) [24].

3. Material and Methods

3.1 Data

The databases were obtained from the Ministry of Higher Education in Colombia, the

Colombian Institute for the Promotion of Higher Education (ICFES - Instituto Colombiano para el Fomento de la Educación Superior) [31] and four (4) private universities of this country. Such data consisted of the reports described in Table 1.

Table 1. Database of the study sample.

DataBase	Description
Student	Personal data of students and their status.
Subject	Data of the subjects taught, and entry conditions of universities under study.
Region	Regions and cities where students come from.
Opportunity	Data on possible opportunities to study the subjects.
Advance time	Permanence time of a student in the career, based on semesters.
Geographical Area	Geographical area where the student is located.
Cohort	Cohort to which students belong.

3.2 Methods

3.2.1 Implementation of the Data Warehouse

A DW system can be implemented under Molap approach (MultidimensionalOlap), Rolap (RelacionalOlap) or by using the hybrid Holap (allows both Molap and Rolap) [32]. In this study, Rolap approach was used. Independently from the approach, the main processes carried out in the development of a DW are as follows.

The process of conceptual modeling: The conceptual model is independent from technology and is essential for specifying the analysis requirements and information availability. When talking about DW conceptual models, there is no consensus in the scientific community about a standard model type for the representation of a DW. However, there are various proposals presented in [33], [34], [35]. During the process of conceptual modeling, a DW conceptual scheme is generated. In this study, the MCMD conceptual model was used [3] due to its notation simplicity and because its objective is precisely the conceptual specification of a DW.

Logical modeling process and physical implementation: The logical model formally specifies the multidimensional scheme, its restrictions, and capabilities. In the same way, the logical scheme is implemented directly in a database engine, becoming physical tables. In the case of DW schemes with logical design, they are the star scheme and snowflake scheme [32]. At the stage of physical implementation, dimension tables and fact tables are created depending on the type of scheme, whether star or snowflake.

ETL data load process: The ETL (Extraction, Transformation, Load) process is responsible for extracting, transforming, and loading the data from the original databases into the DW. The data retrieval approach is proposed from the Intelligent Distributed Data Warehouse (IDDW), which is a hierarchical distributed data store of N levels. Based on Abhay K. et al. (2017) [36], the data retrieval approach begins when the user enters the UIN (Unique Identification Number) corresponding to the data store located in IDDW. Once the data store is located, the desired data are retrieved. A flowchart of the IDDW data retrieval approach is shown in Figure 1.

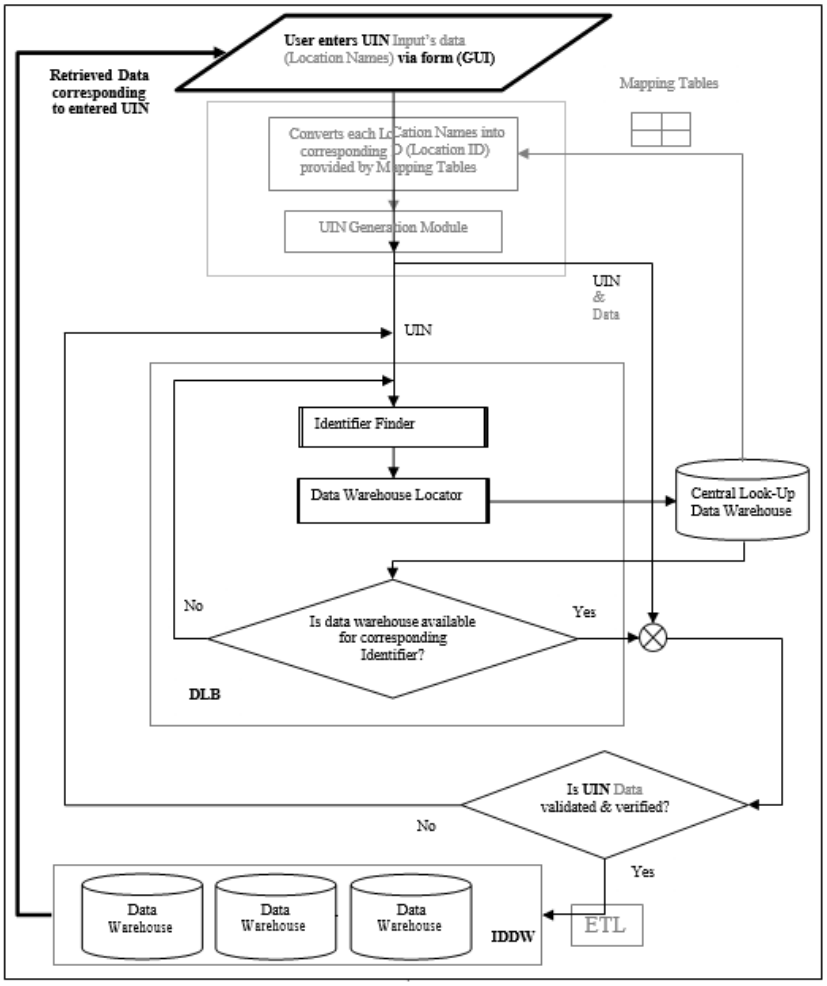

Fig. 1. Flowchart of the data retrieval approach from the IDDW [36].

The ETL process in Figure 1 consists of extracting data from the system database of the university student´s curriculum information, which is not supported by a relational engine and works through files (legacy systems). This system is accessible only through a user interface over the network via a console application inherited

from the COBOL language. To remove this information, the manual process of extraction was simulated by means of an application specially designed for this purpose, after which, the curriculum of each student was extracted in text format. These text files were transformed using a custom software and loaded to a relational database. Then, the files were transformed again by another application for loading them to the DW.

3.2.2 Implementation of the ANN architecture.

At this stage, the ANN architecture was created to be fed with some data obtained by means of the DW. After uploading the DW, an ANN architecture was designed for predicting student´s performance using MATLAB algorithms. In this case, the ANN was used to estimate the behavior of a student in next semester. The neural network was trained with backpropagation algorithm and the sigmoid logarithmic function was used on both layers of the network [7].

The obtained results were validated using performance measures that indicate the generalization degree of the used model. Among the indices used are [8]: The Mean Square Error (MSE), the Residual Standard Error (RSE) and the Index of Adequacy (IA), shown in equations (1), (2), and (3) respectively, where oi and pi are observed and predicted values respectively, in the time i, and N is the total number of data. In addition, pi'= pi - om and oi'= oi - om, om representing the average value of the observations.

The IA indicates the adjustment degree that the estimated values present with the actual values of a variable. A value close to 1 indicates a good estimate. On the other hand, MSE and RSE close to zero indicate a good adjustment quality [8].

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n} o_i^2}} \quad \textbf{(1)}$$

$$RSD = \sqrt{\frac{\sum_{i=1}^{n}(o_i - p_i)^2}{N}} \quad \textbf{(2)}$$

$$IA = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(|o_i'| + |p_i'|)^2} \quad \textbf{(3)}$$

4. Analysis and Results

This section analyzes the behavior of certain indicators over time through the DW architecture implemented and the prediction of any of these indicators through an ANN.

4.1 DW analysis

In order to validate the IDDW operation when integrating the generated profiles, 1.500 queries were carried out, with a limit of 860 records to be retrieved for each by executing the data retrieval processes mentioned in Figure 1. The effectiveness of the IDDW was evaluated in three aspects: (a) the storage of the links to the profiles, (b) the retrieval of the entity data, and c) the registry of the relationships between the entities retrieved from the same document [36]. The results obtained can be seen in Table 2.

Table 2. Validation Results

Metrics	Value
Number of profiles to generate	750
Effectiveness of persistence	90%
Effectiveness of retrieval	76%
Effectiveness of the relationship generation between entities	95%

Initially, the UIN "13302010410520017" is entered through the developed form (Abhay K. et al., 2017) [36]. The first identifier calculated by the identifier search engine for this UIN is "1330201041052001". The data store locator searches for the address of the machine, corresponding to this identifier in the Central Look-Up data store tables. For levels of hierarchy see Table 3.

Table 3. The percentage of correctly retrieved data from the Common table of the data store located at different levels of hierarchy.

Level in hierarchy	Percentage (%)
1	90
2	91
3	93
4	94
5	95
6	98
7	99

Table 3 shows the correctly retrieved data (in percentage) from the data warehouse located in various levels of hierarchy. The correct data are the data that must be retrieved for the entered UIN. From the values in Table 2, it may be seen that as the data warehouse is placed at lower levels of hierarchy, the percentage of correct data retrieved increases. It is because the number of times the Identifier is calculated are less, and chances of error are less too.

4.2 Results of the prediction using ANN.

The number of subjects enrolled by a student was estimated, analyzing the 760 recovered data, taking only one semester toward the future (number of courses approved). With respect to the foregoing, Table 4 shows the values of the indices obtained for estimating both variables.

Table 4. Indices of adequacy and errors in test data estimation.

Indices	Estimation of quantity of enrolled subjects	Estimation of quantity of approved subjects
IA	0.8714	0.8124
RMS	0.3001	0.3492
RSD	**0.0899**	**0.1199**

The results confirm that the prediction is adjusted to the DW historical trend. So, the complement between DW and ANN is a powerful tool to predict the future behavior of a management indicator.

5. Conclusions

The implementation of a Data Warehouse and the Artificial Neural Network architecture has been carried out for the analysis and prediction of academic performance in students of Industrial Engineering at a group of Colombian private universities. The main advantage of using a DW lies in the possibility of crossing different analysis dimensions in a simple and fast way to perform an exploratory analysis of data for the creation of reports. It can be noted that the process of ETL (Extraction, Transformation, and Loading) is the one that more time and resources demanded, mainly since the information should be cross-posted from different sources. Additionally, operational systems are not designed to analyze data, and the heterogeneity of the platforms where the information is located adds a greater difficulty that requires the creation of specific applications and systems to draw on historical data. The use of a multidimensional conceptual model to generate the IDDW conceptual scheme with UIN becomes a great tool that, independently from the platforms, allows to narrow down the analysis and give clarity to the ETL subsequent process.

To obtain summaries and reports using DW as a product of the historical analysis of data, a solid database can be created for the ANN architecture and the prediction of future behavior. Based on the above, the use of DW combined with the use of estimation or prediction techniques (in our case, the ANN), provides a complement to substantiate more extensive analyzes because, as shown in this study, it is possible to predict the management indicators obtained from the DW. This allows the institution to take steps to analyze, modify, and validate the management indicators or, perhaps, to generate new strategies to improve and/or optimize the management process, since

knowledge is extracted from the same databases, thus giving value to the management information that is logged but that is not always considered.

References

1. Vasquez, C., Torres, M., Viloria, A.: Public policies in science and technology in Latin American countries with universities in the top 100 of web ranking. J. Eng. Appl. Sci. **12**(11), 2963–2965 (2017).
2. Aguado-López, E., Rogel-Salazar, R., Becerril-García, A., Baca-Zapata, G.: Presencia de universidades en la Red: La brecha digital entre Estados Unidos y el resto del mundo. Revista de Universidad y Sociedad del Conocimiento **6**(1), 1–17 (2009).
3. Torres-Samuel, M., Vásquez, C., Viloria, A., Lis-Gutiérrez, J.P., Borrero, T.C., Varela, N.: Web Visibility Profiles of Top100 Latin American Universities. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham, vol **10943**, 1-12 (2018).
4. Viloria, A., Lis-Gutiérrez, J.P., Gaitán-Angulo, M., Godoy, A.R.M., Moreno, G.C., Kamatkar, S.J. : Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching – Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham, vol **10943**,1-12 (2018).
5. Caicedo, E.J.C., Guerrero, S., López, D.: Propuesta para la construcción de un índice socioeconómico para los estudiantes que presentan las pruebas Saber Pro. Comunicaciones en Estadística, vol. **9**(1), 93-106 (2016).
6. Mazón, J.N., Trujillo, J., Serrano, M., Piattini, M.: Designing Data Warehouses: From Business Requirement Analysis to Multidimensional Modeling. In Proceedings of the 1st Int. Workshop on Requirements Engineering for Business Need and IT Alignment. Paris, France (2005).
7. Vásquez, C., Torres-Samuel, M., Viloria, A., Lis-Gutiérrez, J.P., Crissien Borrero, T., Varela, N., Cabrera, D.: Cluster of the Latin American Universities Top100 According to Webometrics 2017. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham , vol **10943**, 1-12 (2018).
8. Haykin, S.: Neural Networks a Comprehensive Foundation. Second Edition. Macmillan College Publishing, Inc. USA. ISBN 9780023527616 (1999).
9. Isasi, P., Galván, I.: Redes de Neuronas Artificiales. Un enfoque Práctico. Pearson. ISBN 8420540250 (2004).
10. Haykin, S.: Neural Networks and Learning Machines. New Jersey, Prentice Hall International (2009).
11. Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing **50** (1), 159-75 (2003).
12. Kuan, C.M.: Artificial neural networks. In the New Palgrave Dictionary of Economics, ed. S.N. Durlauf and L.E Blume. UK: Palgrave Macmillan (2008).
13. Jain, A. K., Mao, J., Mohiuddin, K. M.: Artificial neural networks: a tutorial. IEEE Computer **29** (3), 1- 32 (1996).
14. Sevim, C., Oztekin, A., Bali, O., Gumus, S., Guresen, E.: Developing an early warning system to predict currency crises. European Journal of Operational Research **237(**1), 1095-104 (2014).
15. Sekmen, F., Kurkcu, M.: An Early Warning System for Turkey: The Forecasting of Economic Crisis by Using the Artificial Neural Networks. Asian Economic and Financial Review **4**(1), 529-43 (2014).
16. Singhal, D., Swarup, K.S.: Electricity price forecasting using artificial neural networks.

IJEPE **33** (1), 550-55 (2011).
17. Mombeini, H., Yazdani-Chamzini, A.: Modelling Gold Price via Artificial Neural Network. Journal of Economics, Business and Management **3** (7), 699-703 (2015).
18. Kulkarni, S., Haidar, I.: Forecasting Model for Crude Oil Price Using Artificial Neural Networks and Commodity Future Prices. International Journal of Computer Science and Information Security **2** (1), 81-89 (2009).
19. Bontempi, G., Ben Taieb, S., Borgne, Y. A.: Machine learning strategies for time series forecasting. In Lecture Notes in Business Information Processing, ed M.-A. Aufaure., and E. Zimányi, Heidelberg: Springer **138** (1), 70-73 (2013).
20. Duan, L., Xu, L., Liu, Y., Lee, J.: Cluster-based outlier detection. Annals of Operations Research **168** (1), 151–168 (2009).
21. Abhay, K. A., Badal, N. A.: Novel Approach for Intelligent Distribution of Data Warehouses. Published in Egyptian Informatics Journal-Elsevier, Egypt **17** (1), 147-159, (October, 2015).
22. Savasere, A., Omiecinski, E., Navathe, S.: An efficient algorithm for data mining association rules in large databases", In Proceedings of 21st Very Large Data Base Conference, **5** (1), 432- 444 (1995).
23. Stolfo, S., Prodromidis, A. L., Tselepis, S., Lee, W., Fan, D. W.: Java agents for metalearning over distributed databases. In Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining **5** (2), 74-81 (1997).
24. Prodromidis, A., Chan, P. K., Stolfo, S. J.: Meta learning in distributed data mining systems: Issues and approaches. In Kargupta H., Chan P. (eds) Book on Advances in Distributed and Parallel Knowledge Discovery, AAAI/MIT Press (2000).
25. Parthasarathy, S., Zaki, M.J., Ogihara, M.: Parallel data mining for association rules on shared-memory systems, Knowledge and Information Systems: An International Journal **3**(1), 1-29, (February, 2001).
26. Grossman, R. L., Bailey, S. M., Sivakumar, H. Turinsky, A. L.: Papyrus: a system for data mining over local and wide area clusters and super-clusters. In Proceedings of ACM/IEEE Conference on Supercomputing, Article **63**, 1-14, (1999).
27. Chattratichat, J., Darlington, J., Guo, Y., Hedvall, S., Kohler, M. Syed, J.: An architecture for distributed enterprise data mining. In Proceedings of 7th International Conference on HighPerformance Computing and Networking, Netherlands, April 12–14, 573-582 (1999).
28. Wang, L., Tao, J., Ranjan, R., Marten, H., Streit, A., Chen, J., Chen, D.: G-Hadoop: MapReduce across distributed data centers for data-intensive computing. Future Generation Computer Systems **29**(3), 739-750 (2013).
29. Butenhof, D. R.: Programming with POSIX threads. Addison-Wesley Longman Publishing Company, USA (1997).
30. Bhaduri, K., Wolf, R., Giannella, C., Kargupta, H.: Distributed decision-tree induction in peer-to-peer systems. Statistical Analysis and Data Mining **1** (2), 85-103 (2008).
31. Instituto colombiano para la Evaluación de la Educación - ICFES. Informe nacional de resultados Saber Pro 2015–2018. Bogotá: ICFES (2018).
32. Rafailidis, D., Kefalas, P., Manolopoulos, Y.: Preference dynamics with multimodal user-item interactions in social media recommendation. Expert Systems with Applications **74**(1), 11-18 (2017).
33. Zheng, C., Haihong, E., Song, M., Song, J.: CMPTF: Contextual Modeling Probabilistic Tensor Factorization for recommender systems. Neurocomputing **205**(1), 141-151 (2016).
34. Hidasi, B., Tikk, D.: Fast ALS-based tensor factorization for context-aware recommendation from implicit feedback. Machine Learning and Knowledge Discovery in Databases (2012).
35. Lee, J., Lee, D., Lee, Y. C., Hwang, W. S., Kim, S. W.: Improving the accuracy of top-n recommendation using a preference model. Information Sciences **348**(1), 290-304 (2016).
36. Abhay, K.A., Neelendra, B.: Data Storing in Intelligent and Distributed Data Warehouse using Unique Identification Number, published in International Journal of Grid and

Distributed Computing, Publisher: SERSC Australia **10**(9), 13-32 (September, 2017).